Escola de
ECONOMIA
de São Paulo

Textos para Discussão

163

Julho
de 2007

CORRECTING THE FIXED-EFFECT ESTIMATOR FOR ENDOGENOUS SWITCHING

FERNANDO B. BOTELHO AND VLADIMIR P. PONCZEK

FGV
EESP

Correcting the fixed-effect estimator for endogenous switching

Fernando B. Botelho and Vladimir P. Ponczek

RESUMO

PALAVRAS CHAVES

CLASSIFICAÇÃO JEL

C33; C35; C51

ABSTRACT

In this paper, we propose a two-step estimator for panel data models in which a binary covariate is endogenous. In the first stage, a random-effects probit model is estimated, having the endogenous variable as the left-hand side variable. Correction terms are then constructed and included in the main regression.

KEY WORDS

Panel data, Fixed-effect estimator, Endogenous switching

Os artigos dos *Textos para Discussão da Escola de Economia de São Paulo da Fundação Getulio Vargas* são de inteira responsabilidade dos autores e não refletem necessariamente a opinião da FGV-EESP. É permitida a reprodução total ou parcial dos artigos, desde que creditada a fonte.

Escola de Economia de São Paulo da Fundação Getulio Vargas FGV-EESP
www.fgvsp.br/economia

Correcting the fixed-effect estimator for endogenous switching

Fernando B. Botelho and Vladimir P. Ponczek

May 10, 2007

The most important advantages of the fixed-effect estimator for panel data models is the ability to control for unobservable attributes which are constant over time. In most cases, it delivers unbiased estimates and is easily implemented. Nevertheless, it is still possible to have unobservable variables that are not constant over time but are corretated with another covariate. In this scenario, the fixed-effect estimator will not deliver unbiased estimators.

We propose a two-step estimator for panel data models in which a binary covariate is endogenous. In the first stage, a random-effects probit model is estimated, having the endogenous variable as the left-hand side variable. Correction terms are then constructed and included in the main regression.

The literature has focused on the problem related to the estimation of panel models with selectivity bias. [**?**] develop a test to check the presence of selectivity bias based on Heckman-type tests. [**?**] present two-step estimators for a range of parametric panel models, which encompass the problems addressed by [**?**]. Different from the previous cited works that rely on strong assumptions of the error and individual fixed-effect terms, [**?**] relaxes these assumptions. She follows [**?**] procedures and proposes a two-step semi-parametric estimator, which 'differences out' both individual fixed-effect and sample selection bias. [**?**] has an excellent survey about the sample selection literature, including panel data models. Our paper contributes to the literature by adapting the two-step parametric procedures that deal with selectivity bias to the endogenous switching problem. In this case, the dichotomic endogenous variable is ruled by a choice equation and it is potentially correlated to unobserved characteristics of the individual that are not fixed.

One motivation for our estimator is the measurement the wage differential between formal and informal jobs. Our main objetive is to estimate the effect of the job status on the hourly earnings. The ordinary fixed-effect estimator can deliver unbiased estimates if the choice between formal and informal jobs is determined by some characteristic intrinsic to the worker and constant over time. Otherwise,

Suppose the variables w_{it} and b_{it} are determined by following equations:

$$w_{it} = \theta b_{it} + \beta' X_{it} + \mu_i + \epsilon_{it} \tag{1}$$

$$b_{it} = 1_{(\gamma' Z_{it} + \alpha_i + \upsilon_{it} \geq 0).} \tag{2}$$

μ_i and α_i are time-invariant individual effects (possibly correlated with each other), and ϵ_{it} and υ_{it} are pure error terms, possibly correlated with each other and with the individual effects. Assume that X_{it} is a vector exogenous variables, and Z_{it} is predetermined. In this case

$$E\left(w_{it} \mid \mathbf{X}_i, \mathbf{Z}_i, \mathbf{b}_i\right) = \theta b_{it} + \beta' X_{it} + E\left(\mu_i \mid \mathbf{X}_i, \mathbf{Z}_i, \mathbf{b}_i\right) + E\left(\epsilon_{it} \mid \mathbf{X}_i, \mathbf{Z}_i, \mathbf{b}_i\right)$$

Demeaning (**??**) to eliminate the individual fixed effect, we obtain

$$E\left(\tilde{w}_{it} \mid \mathbf{X}_i, \mathbf{Z}_i, \mathbf{b}_i\right) = \theta\ \tilde{b}_{it} + \beta'\ \tilde{X}_{it} + E\left(\tilde{\epsilon}_{it} \mid \mathbf{X}_i, \mathbf{Z}_i, \mathbf{b}_i\right),$$

where $\tilde{a}_{it} = a_{it} - \sum_{t=1}^{T} a_{it}/T$, and $\mathbf{a}_i = (a_{i1}, (...), a_{iT})$.

It is easy to see that the fixed effect estimator is consistent if

$$E\left(\tilde{\epsilon}_{it} \mid \mathbf{X}_i, \mathbf{Z}_i, \mathbf{b}_i\right) = 0$$

This would be true if ϵ_{it} and υ_{it} were uncorrelated. Otherwise, the fixed-effect estimator will not deliver an unbiased estimate of the degree of segmentation.

In order to deal with that potential source of bias, we propose a correction term to be added to the main regression such that the fixed-effect estimator will be shielded from the potential selection bias. Define $\mathbf{u}_i = \alpha_i\ \iota + \boldsymbol{\upsilon}_i$, where ι is the unitary column vector of appropriate dimension. To construct that estimator we need to impose the following structure to the model:

$$\begin{bmatrix} \boldsymbol{\epsilon}_i \\ \mu_i \iota \\ \alpha_i \iota + \boldsymbol{\upsilon}_i \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \sigma_\epsilon^2\ \mathbf{I} & \mathbf{0} & \rho_{\epsilon\upsilon}\ \sigma_\eta\ \sigma_\upsilon\ \mathbf{I} \\ \mathbf{0} & \sigma_\mu\ \iota\ \iota' & \rho_{\mu\alpha}\ \sigma_\mu\ \sigma_\alpha\ \mathbf{I} \\ \rho_{\eta\upsilon}\ \sigma_\epsilon\ \sigma_\upsilon\ \mathbf{I} & \rho_{\mu\alpha}\ \sigma_\mu\ \sigma_\alpha\ \mathbf{I} & \sigma_\alpha^2\ \iota\ \iota' + \sigma_\upsilon^2\ \mathbf{I} \end{bmatrix} \right)$$

Under these assumptions, it is straightforward to show that

$$E\left(\epsilon_i \mid \mathbf{u}_i\right) = \rho_{\epsilon\upsilon}\ \sigma_\epsilon\ \sigma_\upsilon\ \left[\sigma_\alpha^2\ \iota\ \iota' + \sigma_\upsilon^2\ \mathbf{I}\right]^{-1} \mathbf{u}_i = \frac{\rho_{\epsilon\upsilon}\ \sigma_\epsilon\ \sigma_\upsilon}{\sigma_\upsilon^2}\left[\mathbf{I} - \frac{\sigma_\alpha^2}{\sigma_\upsilon^2 + T\ \sigma_\alpha^2}\iota\iota'\right]\mathbf{u}_i$$

Each element of this vector has the form

$$E\left(\epsilon_{it} \mid \mathbf{u}_i\right) = \frac{\rho_{\epsilon\upsilon}\ \sigma_\epsilon\ \sigma_\upsilon}{\sigma_\upsilon^2}\left[u_{it} - \frac{T\ \sigma_\alpha^2}{\sigma_\upsilon^2 + T\ \sigma_\alpha^2}\ \frac{\sum_{s=1}^{T} u_{is}}{T}\right].$$

Now, conditioning on the appropriate interval, we obtain

$$E\left(\epsilon_{it} \mid \mathbf{b}_i, \mathbf{Z}_i\right) = \frac{\rho_{\epsilon\upsilon}\ \sigma_\epsilon\ \sigma_\upsilon}{\sigma_\upsilon^2}\left[E\left(u_{it} \mid \mathbf{b}_i, \mathbf{Z}_i\right) - \frac{T\ \sigma_\alpha^2}{\sigma_\upsilon^2 + T\ \sigma_\alpha^2}\ \frac{\sum_{s=1}^{T} E\left(u_{is} \mid \mathbf{b}_i, \mathbf{Z}_i\right)}{T}\right]$$

Finally, by demeaning the previous equation we have

$$\begin{aligned} E\left(\tilde{\epsilon}_i \mid \mathbf{b}_i, \mathbf{Z}_i\right) &= \frac{\rho_{\epsilon\upsilon}\ \sigma_\epsilon\ \sigma_\upsilon}{\sigma_\upsilon^2}\ E\left(\tilde{u}_{it} \mid \mathbf{b}_i, \mathbf{Z}_i\right) \\ &= \frac{\rho_{\epsilon\upsilon}\ \sigma_\epsilon\ \sigma_\upsilon}{\sigma_\upsilon^2}\ E\left(\tilde{\upsilon}_{it} \mid \mathbf{b}_i, \mathbf{Z}_i\right) \end{aligned}$$

This difficulty of this procedure is to calculate $E\left(\upsilon_{it} \mid \mathbf{b}_i, \mathbf{Z}_i\right)$, since b_i is depend of α_i and the vector υ_i. Therefore, one would have to integrate a bivariate normal distribution over the range defined by b_i. However, to avoid this numericaly cumbersome procedure, we observe that

$$E\left(\upsilon_{it} \mid \mathbf{b}_i, \mathbf{Z}_i\right) = \int E\left(\upsilon_{it} \mid \mathbf{b}_i, \mathbf{Z}_i, \alpha_i\right) f\left(\alpha_i \mid \mathbf{b}_i, \mathbf{Z}_i\right) d\alpha_i.$$

It can be seen that

$$E\left(\upsilon_{it} \mid \mathbf{b}_i, \mathbf{Z}_i, \alpha_i\right) = E\left(\upsilon_{it} \mid b_{it}, Z_{it}, \alpha_i\right) = \frac{(2\ b_{it} - 1)\ \phi(\gamma' Z_{it} + \alpha_i)}{\Phi[(2\ b_{it} - 1)(\gamma' Z_{it} + \alpha_i)]},$$

and the Bayes rule implies

$$f\left(\alpha_i \mid \mathbf{b}_i, \mathbf{Z}_i\right) = \frac{f\left(\mathbf{b}_i \mid \mathbf{Z}_i, \alpha_i\right) f\left(\alpha_i \mid \mathbf{Z}_i\right)}{\int f\left(\mathbf{b}_i \mid \mathbf{Z}_i, \alpha_i'\right) f\left(\alpha_i' \mid \mathbf{Z}_i\right) d\alpha_i'}$$

One of the advantages of this procedure is to allow us to perform two one-dimensional instead of one two-dimensional numerical integrations.

Our method consists in estimating the model in two steps. First, we have to estimate γ and σ_{alpha} consistently. This is done by a random effect probit in the selection equation. Notice that an important for the consistency of those estimators in the first step is

$$f\left(\alpha_i \mid \mathbf{Z}_i\right) = f\left(\alpha_i\right).$$

In words, it means that fixed effect in the selection equation (α_i) is orthogonal to the vector of instruments ($\mathbf{Z}_i$). Therefore,

$$\frac{f\left(\mathbf{b}_i \mid \mathbf{Z}_i, \alpha_i\right) f\left(\alpha_i\right)}{\int f\left(\mathbf{b}_i \mid \mathbf{Z}_i, \alpha_i'\right) f\left(\alpha_i'\right) d\alpha_i'} = \frac{\Phi[(2\ b_{it}-1)(\gamma' Z_{it}+\alpha_i)]\frac{1}{\sqrt{2\pi\sigma_\alpha^2}}\exp\left(-\frac{1}{2}\left(\frac{\alpha_i}{\sigma_\alpha}\right)^2\right)}{\int \Phi[(2\ b_{it}-1)(\gamma' Z_{it}+\alpha_i')]\frac{1}{\sqrt{2\pi\sigma_\alpha^2}}\exp\left(-\frac{1}{2}\left(\frac{\alpha_i'}{\sigma_\alpha}\right)^2\right) d\alpha_i'},$$

so the correction term is given by

$$E\left(\upsilon_{it} \mid \mathbf{b}_i, \mathbf{Z}_i\right) = \int \frac{(2\ b_{it}-1)\ \phi(\gamma' Z_{it}+\alpha_i)}{\Phi[(2\ b_{it}-1)(\gamma' Z_{it}+\alpha_i)]} \times \frac{\Phi[(2\ b_{it}-1)(\gamma' Z_{it}+\alpha_i)]\frac{1}{\sqrt{2\pi\sigma_\alpha^2}}\exp\left(-\frac{1}{2}\left(\frac{\alpha_i}{\sigma_\alpha}\right)^2\right)}{\int \Phi[(2\ b_{it}-1)(\gamma' Z_{it}+\alpha_i')]\frac{1}{\sqrt{2\pi\sigma_\alpha^2}}\exp\left(-\frac{1}{2}\left(\frac{\alpha_i'}{\sigma_\alpha}\right)^2\right) d$$

In order to have a standard normal distribution of the fixed effect, we have to redefine the integration variable as

$$r_i = \frac{\alpha_i}{\sqrt{2\sigma_\alpha^2}} \rightarrow \alpha_i = s\ r_i$$

Finally, the correction term can be written as

$$\begin{aligned} E\left(\upsilon_{it} \mid \mathbf{b}_i, \mathbf{Z}_i\right) &= \int \frac{(2\ b_{it}-1)\ \phi(\gamma' Z_{it}+s\ r_i)}{\Phi[(2\ b_{it}-1)(\gamma' Z_{it}+s\ r_i)]} \times \frac{\Phi[(2\ b_{it}-1)(\gamma' Z_{it}+s\ r_i)]\exp\left(-r_i^2\right)}{\int \Phi[(2\ b_{it}-1)(\gamma' Z_{it}+s\ r_i')]\exp\left(-r_i'^2\right) dr_i'} dr_i \\ &\approx \sum_{r_i \in R} \frac{(2\ b_{it}-1)\ \phi(\gamma' Z_{it}+s\ r_i)}{\Phi[(2\ b_{it}-1)(\gamma' Z_{it}+s\ r_i)]} \times \frac{\Phi[(2\ b_{it}-1)(\gamma' Z_{it}+s\ r_i)]\exp\left(-r_i^2\right)}{\sum_{r_i' \in R}\Phi[(2\ b_{it}-1)(\gamma' Z_{it}+s\ r_i')]\exp\left(-r_i'^2\right)} \end{aligned}$$

The last term is a numerical approximation obtained by the Gauss-Hermite numeric integration method.

We then introduce the correction term into the main equation to eliminate the possible endogeneity of b_{it}. Let us define $corr_i$ as the vector with correction terms and $W_i = [b_i \ X_i \ corr_i]$. In vector notation, equation **??** augmented by the correction term becomes:

$$w_i = \phi' W_i + \mu_i + \xi i, \tag{3}$$

where $\phi = (\theta, \beta, \lambda)$. Also we define the following terms[1]:

$$M = N^{-1} \sum_{i=1}^{N} E[W_i' W_i]$$

$$V = N^{-1} \sum_{i=1}^{N} E[W_i' Var(\xi) W_i]$$

$$D = N^{-1} \sum_{i=1}^{N} E[W_i' \frac{\partial \lambda(\gamma)}{\partial \gamma}]$$

ϕ is asymptotically normally distributed with covariance matrix[2]:

$$\lim_{n \to \infty} M^{-1}(V + DHD')M^{-1} \tag{4}$$

where H is the Hessian matrix generated in the maximum likelihood estimation in the first stage.

In the absence of the endogeneity problem, λ is zero and DHD' is also zero, thus, the covariance matrix becomes trivial. A straightforward way to check whether $b_{i,t}$ is endogenous is to test the statistical significance of λ.

[1] We are assuming that ξ_i has spherical variance.

[2] Following [**?**]